DOCUMENT RESUME

ED 320 921 TM 014 990

AUTHOR Englehard, George, Jr.
TITLE Thorndike, Thurstone and Rasch: A Comparison of Their Approaches to Item-Invariant Measurement.
SPONS AGENCY National Academy of Education, Washington, D.C.
PUB DATE 2 May 90
NOTE 52p.; Paper presented at the Annual Meeting of the American Educational Research Association (Boston, MA, April 16-20, 1990). Table 1 contains small type.
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS Ability Identification; Comparative Analysis; Elementary School Students; Estimation (Mathematics); Grade 5; Intermediate Grades; Item Response Theory; Performance; *Scoring; *Test Items; *Theory Practice Relationship
IDENTIFIERS Ability Estimates; *Item Invariance

ABSTRACT

The methods used by E. L. Thorndike, L. L. Thurstone, and G. Rasch to address issues related to item-invariant measurement and the scoring of individual performance are compared. The analyses highlight the close connection among the three methods, and suggest that progress in measurement theory reflects the movement from essentially ad hoc methods to theory-based procedures for dealing with measurement problems. Data from a 1914 study by Thorndike involving 40 fifth-grade students are used to illustrate how Thorndike, Thurstone, and Rasch would approach the problem of obtaining item-invariant estimates of ability on a visual vocabulary scale. Five tables and eight figures illustrate the comparisons. A 40-item list of references is included. (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
GEORGE ENGELHARD, JR.
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED320921

THORNDIKE, THURSTONE AND RASCH: A COMPARISON OF THEIR APPROACHES TO ITEM-INVARIANT MEASUREMENT

George Engelhard, Jr.

Emory University

Address: Professor George Engelhard, Jr.
Emory University
Division of Educational Studies
210 Fishburne Building
Atlanta, GA 30322

(404) 727-0607

Acknowledgments

Support for this research was provided through a Spencer Fellowship from the National Academy of Education.

Running head: THORNDIKE, THURSTONE & RASCH

[invari2 - Paper presented as the annual meeting of the American Educational Research Association in Boston, April 1990]

May 2, 1990

TM014990

BEST COPY AVAILABLE

Abstract

The purposes of this study are to describe and compare the methods used by Thorndike, Thurstone and Rasch to address issues related to the item-invariant measurement and the scoring of individual performance. The analyses highlight the close connections between the three methods, and suggest that progress in measurement theory reflects the movement from essentially ad hoc methods to theory-based procedures for dealing with measurement problems. Data from Thorndike (1914) are used to illustrate how Thorndike, Thurstone and Rasch would approach the problem of obtaining item-invariant estimates of ability on a visual vocabulary scale.

THORNDIKE, THURSTONE AND RASCH: A COMPARISON OF THEIR APPROACHES TO ITEM-INVARIANT MEASUREMENT

One of the perennial problems in educational and psychological testing is how to obtain meaningful scores from individual responses to a set of test items. The basic structure of the data being modelled is usually summarized in a matrix (individuals by items) with 0s for wrong responses and 1s for right responses. Although a variety of item response formats are available, the predominant item response format remains the multiple-choice item with dichotomous (0,1) scoring. Given a matrix of dichotomous responses, the measurement problem examined in this paper is how to obtain a score for each individual that is not dependent on the particular items that were administered. The goal of item-invariant measurement is to minimize the effects of the particular items which happen to be used to estimate an individual's ability. Ideally, an individual should obtain a comparable estimate of ability regardless of which items are included in the test.

The general problem of invariance in measurement has been succinctly stated by Wright (1968) in terms of the conditions necessary for objective measurement. In his words,

> First, the calibration of measuring instruments must be independent of those objects that happen to be used for calibration; second, the measurement objects must be independent of the instrument that happens to be used for

> measuring. In practice, these conditions can only be approximated. But this approximation is what makes measurement objective. (Wright, 1968, p. 87)

A historical perspective on general issues related to invariance and measurement theory is provided in Engelhard (1989). Issues related to Wright's first condition for objective measurement related to sample-invariant item calibration within the measurement theories of Thorndike, Thurstone and Rasch are presented in Engelhard (1984).

The focus of this study is on Wright's second condition for objective measurement, and it is referred to here as the goal of item-invariant measurement. Many of the issues related to item-invariance appear within the context of the scoring of individual performance. Thorndike, Thurstone and Rasch were selected because they recognized the importance of item-invariant measurement, and proposed innovative statistical procedures for examining this issue. Another reason for selecting these measurement theorists is that their work illustrates the progress that has taken place related to item-invariant measurement during the twentieth century.

Purpose

The purpose of this paper is to provide a historical perspective on the concept of item-invariant measurement and the scoring of individual performance using the measurement theories of Thorndike, Thurstone and Rasch. The specific questions addressed are as follows:

(1) How are issues related to item-invariant measurement reflected within the theories of Thorndike, Thurstone and Rasch?

(2) What are the similarities and differences in the approaches to item-invariant measurement as conceptualized by Thorndike, Thurstone and Rasch?

In the first section of this paper, the methods recommended by Thorndike, Thurstone and Rasch for addressing issues related to item-invariant measurement, and for scoring individual performance are described. Next, the three methods are compared based on conceptual issues related to item-invariant measurement. Finally an empirical example is presented. The responses of 40 children to a reading scale developed by Thorndike (1914) for measuring visual vocabulary is analyzed with the three methods, and the results compared.

Description of Three Methods

Thorndike

Thorndike wrote extensively on measurement issues in psychology and education (Thorndike, 1904, 1910, 1914, 1918, 1921; Thorndike, Bregman, Cobb, & Woodyard, 1926). Although Thorndike's views on item-invariant measurement are found in several places, I have selected his article on the measurement of ability in reading (Thorndike, 1914) as being representative of his contributions in this area. This article also includes a fairly complete data set that is analyzed in this paper.

Thorndike (1914) presented several scales for measuring school achievement in reading. Thorndike's motivation for developing these scales was to measure how well a child can read with the "objectivity, precision, commensurability and convenience which characterize our measurements of how tall he is" (p. 1). The scale discussed in this paper is called the Thorndike Reading Scale A and is designed to measure visual vocabulary. The responses of 40 fifth-grade students are presented in Table 1. The entries in the Table 1 are the number of

---

Insert Table 1 about here

---

wrongly marked and omitted words for each item group with the correct responses left blank. Thorndike grouped items with comparable scale values, and items with lower scale values are defined as being easier, while items with higher scale values are more difficult. For example, the words "camel" and "kind" are easy and have scale values of 4 on Thorndike's vocabulary scale, and the words "armadillo" and "iguana" are hard and have item scale values of 11. The scale values used by Thorndike with the number of items given in parentheses are as follows: 4.0 (5), 5.0 (5), 6.0 (5), 7.0 (5), 8.0 (5), 9.0 (5), 10.0 (5), 10.5 (5) and 11.0 (3).

Thorndike recommended assigning student scores based on the highest numbered item group with one (or no) omission or error. This

corresponds to a scoring criterion of 80 percent (4 out of 5 items correct for an item group), although Thorndike recognized that this criterion could be replaced by other values such as 50 percent or even 100 percent. In order to obtain an ability estimate, Thorndike summarized the responses to items with the similar scale values. Examples of summarized response patterns are shown in Table 2.

Insert Table 2 about here

In estimating student ability, Thorndike demonstrated his awareness of the problem of item-invariant measurement in several ways. Using student data from 8 classrooms, he points out that "If the measures were precise, the order of the eight classes would be the same by any lines [item scale values] results" (p. 8). Within the framework of this paper, this reflects Thorndike's concern with the invariance of ability estimates over subsets of items of differing difficulty.

Since the scale values for the items are considered known, Thorndike also recognized that individual response patterns can be examined for inconsistent responses for each individual. Inconsistent response patterns imply that student scores will vary depending on which subset of items are used to estimate student ability. For example, Thorndike points out that some students may do better with a more difficult item than with the preceding and easier items. Thorndike

identified several students (for example, students 2, 5, 8, 9 and 18 shown in Table 2) with inconsistent response patterns across items, and made adjustments in the ability estimates of some of these students. He also identified a student (15) with an unscorable response pattern. In the case of inconsistent response patterns, he recommended using the student's "score in the other lines to form a reasonable rough estimate" (p. 10) of ability, and "by a proper set of rules this can be done without personal caprice" (p. 10). The methods for identifying inconsistent response patterns, and his "rules" for adjusting the ability estimate are never explicitly described . The adjustments made for inconsistent responses appear to be early examples of an ad hoc procedure for the robust estimation of ability. It should be stressed that in 1914, most estimates of student ability were based on total test scores, and little if any consideration was given to the reasonableness of an individual's response pattern, or to methods for "adjusting" ability estimates when inconsistent response patterns were identified. Thorndike's students with inconsistent response patterns are examined in more detail later in this paper.

Although Thorndike did not plot a person response curve (Trabin & Weiss, 1983), the grouping of each individual's responses by scale values with comparable difficulties as shown in Table 2 is essentially the same process. In his later work, Thorndike presents "average" person response curves that he recommends using to smooth individual

response patterns. (See Chapter XI in Thorndike, et al., 1926).

In summary, Thorndike (1914) offered what appear to be early examples of a tabular summary of responses that is conceptually related to person response curves, and also person fit analyses with adjustments of ability estimates when inconsistent response patterns are encountered. His uses of these concepts were essentially ad hoc, and not based on any explicit theoretical rationale.

Thurstone

Thurstone's ideas about item-invariant measurement can be viewed in terms of two different approaches. One approach is essentially a group-based approach, while the other offers an individual-level perspective on item-invariance. A group-based approach can be based on the application of his method absolute scaling (Thurstone, 1925, 1927, 1928a, 1928b) to test scores as described by (Gulliksen, 1950). His graphical method for examining the invariance of item scale values over grade groups could also be used to examine invariance of ability estimates over item-score groups. In essence, this analysis would examine whether or not the ordering of individuals is invariant over item-score groups. I have not found any published examples of Thurstone using this approach to explore issues related to item-invariant measurement.

A more interesting treatment of issues related to item-invariance at the individual level of analysis is found in Thurstone (1926), and

therefore I will use this article to represent his work on this measurement problem. In this article, Thurstone explicitly stated a set of requirements for obtaining an item-invariant estimate of an individual's ability. The specific requirements are as follows:

1. It should not be required to have the same number of test elements at each step of the scale.
2. It should be possible to omit several test questions at different levels of the scale without affecting the individual score.
3. It should be possible to include in the same scale two forms of test, namely (a) those that are scored either right or wrong by the all-or-none principle, and (b) those that give a variable score such as time in seconds, or number of right answers or number of errors.
4. It should not be required to submit every subject to the whole range of the scale. The starting point and the terminal point, being selected by the examiner, should not directly affect the individual score.
5. It should be possible to use the scale so that a rational score may be determined for each individual subject and so that the performance of groups of subjects may be compared.
6. The arithmetical labor in determining individual scores should be a minimum.

7. The procedure should be as far as possible consistent with psychophysical methods so that it will be free from the logical errors involved in the Binet scale and its variants.

In order to meet these requirements for scoring individual performance, Thurstone (1926), as did Thorndike (1914), proposed using what would be recognized today as a person response curve. Essentially, Thurstone recommended that items with similar scale values on the latent variable scale be grouped into class intervals or item groups, and the proportion of right answers within each class interval calculated. These proportions can then be used to determine an individual's position on a latent variable scale in ways "analogous to the psychophysical problem of determining a difference limen" (p. 448). According to Thurstone, "the scale value at which the subject gets half the problems right and half of them wrong" (p. 449) can be used as an estimate of the individual's ability and location on the scale. An example of such a empirical person response curve is shown in Figure 1. One problem with this approach, and also with Thorndike's approach, is that an individual may have a response pattern with several item groups that meet the 50 or 80 percent criterion. For the empirical response curve presented in Figure 1, the 50% criterion is met by two item groups that could lead to ability estimates for this individual of 5 or 6 on the scale.

Insert Figure 1 about here

Although Thurstone recognized that a theoretical curve could be used to smooth this empirical curve, he recommended defining an individual's score as the "point on the scale at which the sum of right answers above it equals the sum of wrong answers below it" (p. 450). Specifically, he recommended the following steps. First, calculate and plot the cumulative frequencies of right and wrong scores. This is presented graphically Figure 2.

Insert Figure 2 about here

Second, identify the class interval in which the cumulative frequencies of right and wrong scores cross. In Figure 2, the curves cross between 6 and 7 on the scale. Next, Thurstone recommends calculating the fraction of the class interval as follows:

$$x = \frac{R1 - W1}{(R1 - R2) + (W2 - W1)} \qquad (1)$$

where

R1 = Cumulative frequency of right answers up to and including the class interval where the curves cross,

R2 = Cumulative frequency of right answers up to the class interval where the curves cross,

W1 = Cumulative frequency of wrong answers up to the class interval where the curves cross,

W2 = Cumulative frequency of wrong answers up to and including the class interval where the curves cross,

Finally, this fraction, x, is added to the item scale value at the bottom of the class interval. For this example, the ability estimate is 6.44 [6 + (9-5)/((9-3) + (8-5))].

Although Thurstone's method meets the requirements he set, there are several limitations to his approach. One of the limitations is that Thurstone's method does not explicitly include the item scale values in the computation of an individual's score (Brown, Bartelme & Cox, 1933). Given his work on absolute scaling (Thurstone, 1925) and on psychophysics (Thurstone, 1927), he certainly knew how to do this, but he was more concerned in this article with the development of a "workable compromise which will probably serve all practical purposes" (Thurstone, 1926, p. 457). Another limitation of Thurstone's 1926 approach is that because he did not use a formal statistical model, no standard errors are available for the ability estimates. Thurstone (1926) also did not explicitly address issues related to the consistency of individual response patterns, and methods for obtaining adjusted estimates of individual ability when necessary.

Thurstone did not explicitly develop a method for determining individual scores analogous to his methods used for determining item

scale values. These methods were described by Mosier (1940, 1941), who clearly articulated the duality between the methods used for item calibration and individual measurement. Mosier also explicitly developed the concept of a person response curve. Another practical limitation of Thurstone's method is that if the items are too easy or too hard, the cumulative frequency curves for right and wrong answers may not cross.

In summary, Thurstone sought a method for scoring individual performance that would provide invariant estimates of ability over the number of items (allow omitted responses), item response formats and the particular items administered to the individual. This scoring method should also lead to comparable ability estimates for individuals based on different subsets of items. Further, the scoring method should be easy to use and not norm-referenced. Thurstone's method of scoring individual performance is of historical interest as an early attempt to meet these requirements through the use of a person response curve, but better methods are available today for meeting these requirements. For example, the Rasch model described in the next section meets these requirements and provides additional advantages over Thurstone's method of scoring individual performance. This does not detract from the fact that Thurstone's requirements for item-invariant measurement were quite innovative for his time.

Rasch

Based on psychometric research conducted during the 1950s, Rasch (1961, 1966a, 1966b, 1977, 1980/1960) presented a set of ideas and methods which were described by Loevinger (1965) as a "truly new approach to psychometric problems" (p. 151) which can lead to "nonarbitrary measures" (p. 151). One of the major characteristics of Rasch's approach to measurement was his explicit concern with the development of individual-centered techniques as opposed to the group-based measurement models used within classical test theory. In Rasch's words, "individual-centered statistical techniques require models in which each individual is characterized separately and from which, given adequate data, the individual parameters can be estimated" (1980/1960), p. xx).

Problems related to invariance played an important role in motivating the measurement theory of Rasch. As pointed out by Andrich (1988), Rasch presented "two principles of invariance for making comparisons that in an important sense precede, though inevitably lead to, measurement" (p. 18). Rasch's concept of "specific objectivity" and his principles of comparison form his version of the requirements for invariant measurement (Rasch, 1977). According to Rasch,

> The comparison between two stimuli should be independent of which particular individuals were instrumental for the comparison; and it should also be independent of which

> stimuli within the considered class were or might also have been compared. Symmetrically, a comparison between two individuals should be independent of which particular stimuli within the class considered were instrumental for the comparison; and it should also be independent of which other individuals were also compared, on the same or on some other occasion (Rasch, 1961, pp. 331-332).

It is clear in the second half of this quotation that Rasch recognized the importance of item-invariant measurement of individuals. For Rasch, the objective measurement of ability included both item-invariant methods for scoring individual performance as examined in this paper, as well as a non-normative framework for interpreting the meaning of an individual's ability.

Rasch's approach for examining item-invariant measurement is clearly described in his book on Probabilistic Models for Intelligence and Attainment Tests. The LOG method he proposed consists of several steps. First, Rasch recommended summarizing the individual by item matrix into a raw score by item group matrix. Rasch believed that the ordering of individuals in terms of their abilities should, apart from "random disturbances", follow the ordering of the individuals by raw scores. He also believed that if item-invariant measurement is possible for a particular set of data, then the ability estimates should be invariant over different item groups composed of items which vary in

degree of difficulty; the relative ordering of individuals should remain the same over subsets of items that vary in difficulty. In order to examine item-invariance, Rasch calculated the proportion of correct responses within each cell of the raw score by item group matrix and took the logit of these proportions in order to obtain a linear scale. Next, Rasch recommended plotting the ability estimates obtained separately within each item group against the average ability estimates obtained across item groups. If these plots were linear with slopes close to one, then successful item-invariant measurement has been obtained. In this case, the average ability estimate across item groups can be used as an estimate of an individual's ability. If these plots are not linear, then item-invariant measurement has not been obtained, and the data does not fit the Rasch model.

In order to illustrate how Rasch would explore the issue of item-invariant measurement, the results for two subtests, N and F, from the Danish Military Group Intelligence Test (BPP) which were used by Rasch (1980/1960) are presented in Figure 3. The test data were obtained from

---

Insert Figure 3 about here

1,904 recruits who were tested in September 1953. Panel A illustrates successful item-invariant measurement with ability estimates relatively invariant over item groups, while Panel B provides evidence of

unsuccessful item-invariant measurement as shown by the non-parallel lines with unequal slopes. Since Rasch's work, a number of more formal statistical tests have been proposed for examining the fit of data to the Rasch model (Anderson, 1973; Wright & Masters, 1982; Wright & Panchapakesan, 1969; Wright & Stone, 1979).

In a sense, Rasch's graphical test of model fit can be viewed as a group-level perspective based on raw score groups and item groups. The graphical analyses of fit presented by Rasch (1980/1960) parallel in many ways the method proposed by Thurstone (1925) for examining the fit of data to his method of absolute scaling (Engelhard, 1984, 1989).

For the purposes of this paper, the extensions of the Rasch model to provide micro-level examinations of item-invariance for individual response patterns is of interest. It appears that Rasch did not explicitly develop methods for examining the consistency of individual response patterns, although the seeds of this idea are clearly evident in his writing. The analysis of person fit provides a detailed framework for examining item-invariance for the response pattern of each individual. The extension of the Rasch model to provide micro-level examinations of individual response patterns was suggested by Wright and Panchapakesan (1969). Smith (1985, 1986, 1988) has written extensively on the topic of person fit analysis.

In essence, person fit analysis provides a framework for examining the correspondence between the expected and observed pattern of right

responses with the expected responses based on the ability estimate for each individual. An expected response pattern can be presented as a person response curve that represents the probabilities of succeeding on a set of items as a function of the item scale values and the estimated ability of the individual. For dichotomous data, the person response curve based on the Rasch model has the following form:

$$Pr(X_i = 1 \mid b, d_i) = \frac{\exp(b - d_i)}{1 + \exp(b - d_i)} \qquad (2)$$

This equation represents the probability of an individual with an ability estimate of b, answering item i with difficulty, d_i, correctly ($X_i = 1$). In graphical form, this equation is presented in Figure 4.

Insert Figure 4 about here

This is the expected person response curve for an individual with an ability estimate of .571. Typically, the ability estimate, b, is defined in terms of the location on the latent variable scale where the individual has a probability of .50 of answering correctly an item with that scale value. The method used in this paper for estimating ability is the UCON method (Wright & Stone, 1979).

The congruence between the expected and observed response curves can be summarized in several ways. For the Thorndike data that are analyzed in the next section, person fit statistics are reported. The person fit statistic is based on the cube root transformation of the unweighted mean square residual and is described in Wright & Masters (1982).

In summary, the examination of item-invariance can be approached in at least two ways within the framework of the Rasch model. Rasch's original graphical approach is essentially a group-based method for examining whether or not the ordering of individuals is invariant over subsets of items. More recently, the Rasch model has been extended to provide micro-level analyses of item-invariance for individual response patterns. These micro-level analyses of person fit provide a detailed framework for examining item-invariance for each individual. In essence, a person fit analysis examines the correspondence between the observed and expected pattern of right responses with the expected responses based on the ability estimate for each individual. This correspondence can be presented graphically in terms of observed and expected person response curves, as well as by person fit statistics. The expected person response curve based on the Rasch model provides a theoretical curve to model the empirical person response curves that underlie both Thorndike's and Thurstone's methods of scoring individual performance.

Comparison of Thorndike, Thurstone and Rasch

Conceptual issues

Thorndike, Thurstone and Rasch share a number of ideas and assumptions about measurement in the social sciences. These three measurement theorists were all working within a common research tradition based on scaling theory which has its roots in the 19th century work on psychophysics (Engelhard, 1989). Since they were all working within this research tradition, there is much that is congruent within their measurement work.

In approaching the problem of scoring individual performance, they all recognized the problems inherent in the use of raw scores (summation of correct responses), such as lack of item invariance and unequal units, and avoided using them. Instead, they approached this measurement problem using the concept of a latent variable scale defined by a set of calibrated items with known item scale values. Scoring individual performance was then viewed in terms of the problem of finding an individual's location on this latent variable scale based on his or her particular pattern of item responses; the individual's location on this scale was then used as an ability estimate. Although the details of their methods varied, they all shared the concept of a person response curve that reflects the proportion or probability of answering correctly a set of test items located on a latent variable scale. Further, they all recognized the importance of developing

scoring methods that can lead to item-invariant measurement. The group-level approaches for examining item-invariance are also quite similar.

Although there are clear conceptual similarities between the scoring methods of Thorndike, Thurstone and Rasch, there are also important differences which highlight the progress made in this area of measurement. The major difference between Thorndike and Thurstone as compared to Rasch is that Rasch proposed an explicit statistical model for examining the responses of an individual to a set of test items. There are several major consequences of this difference. First, although Thorndike recognized that inconsistent responses can lead to ability estimates that are not invariant over subsets of items, his "person fit" analyses are essentially ad hoc. By proposing a formal statistical model, Rasch provided a theory-based framework for examining the consistency of individual item response patterns through person fit analysis. Another consequence of having a formal statistical model is that standard errors are available for the Rasch ability estimates.

The scoring methods of Thorndike and Thurstone are essentially deterministic, while the Rasch model provides a probabilistic framework. Although Thorndike used a scoring criterion of 80%, while Thorndike and Thurstone used a criterion of 50%, this difference is essentially trivial. Thorndike's and Thurstone's scoring procedures do not provide a good way for handling randomness and uncertainty in individual response patterns. Thorndike recognized that his scoring rule would not

provide a reasonable estimate of ability for each individual. For example, Thorndike had the problem with his scoring method of how to score the performance of an individual who has two item groups that meet the 80 percent criterion. Since Thorndike did not have a formal model for dealing with randomness within an individual's response pattern, he resorted to "rules of thumb" for adjusting ability estimates. In his 1926 article, Thurstone did not address issues related to inconsistent item response patterns, and methods for adjusting ability estimates when inconsistent patterns are evident. Rasch's model for person response curves provides a theory-based approach for estimating ability, and examining item invariance through person fit analysis. Rasch did not directly address the issue of what to do when item-invariance is not obtained. Wainer and Wright (1980) describe several robust estimators of ability based on the Rasch model that offer promising approaches for adjusting ability estimates when inconsistent responses due to guessing and other disturbances occur.

In developing his model, Rasch treated item calibration and individual measurement as simultaneous activities, while Thorndike and Thurstone generally treated them separately. In fact, Thorndike's and Thurstone's views on these two activities are generally presented in separate articles and chapters. For example, relatively little detail on item calibration is presented in either of the major articles used to represent Thorndike's (1914) or Thurstone's (1926) methods for scoring

individual performance. This issue is not as important for the purposes of this paper, since the item scale values are assumed to be known, but it does reflect an important historical milestone in measurement theory. These comparisons of Thorndike, Thurstone and Rasch on these issues are summarized in Table 3.

Insert Table 3 about here

Empirical example

The implications of the different methods for scoring individual performance can be illustrated by analyzing a common data set with all three methods. The data used in this numerical example are from Thorndike (1914) and are shown in Table 1. This data set was described earlier in the Thorndike section.

Before using the three methods to analyze this data, several steps were taken to validate the item scale values. First, Thorndike's item scale values were centered to have a mean of 0. Next, the 43 vocabulary items were calibrated with the Rasch model, and seven items were deleted because all of the 40 fifth-grade students in this sample answered them correctly (items 1, 2, 4, 9, 11, 13 & 15). The average of the Rasch item difficulties were then calculated for each item group used by Thorndike, and assigned to items within that item group. The item scale values obtained by Thorndike and the averaged Rasch item scale values

are virtually identical, $r(41) = .99$, $p < .01$. This close correspondence between the item calibration methods of Thorndike and Rasch has been found with other data sets (Engelhard, 1984). This analysis supports the use of the Thorndike item scale values ($M = 0$, $SD = 2.34$) for the comparisons of ability estimates used throughout the paper. The centered values with the number of items in parentheses are as follows: -3.686 (5), -2.686 (5), -1.686 (5), -.686 (5), .314 (5), 1.314 (5), 2.314 (5), 2.814 (5) and 3.314 (3).

The original Thorndike ability estimates were replaced with estimates based on the centered item scale values. A constant of 1.39 was also added to the Thorndike estimates to adjust for the difference between his criterion of 80 percent and the 50 percent criterion used by Thurstone and Rasch. The scoring methods described earlier for Thurstone was used to obtain the Thurstone ability estimates, while the UCON method was used to obtain the Rasch ability estimates.

The ability estimates obtained by each method are shown in Table 4.

Insert Table 4 about here

There are several notable features of this table. First, the means of the ability distributions for the Thorndike ($M = 2.319$), Thurstone ($M = 2.559$) and Rasch ($M = 2.277$) are very similar. The standard deviations for the Thorndike ($SD = 1.264$) and Rasch ($SD = 1.076$) ability estimates

are comparable, while the standard deviation for the Thurstone ability estimates (SD = .741) is lower. This smaller standard deviation is partly due to the fact that 2 high scoring students (10 and 25) have non-crossing right and wrong cumulative frequency curves, and therefore Thurstone ability estimates are not available.

Another notable feature of Table 4 is that the overall correspondence between the ability estimates obtained by the three different scoring methods is quite high. The highest agreement is found between the Thurstone and Rasch ability estimates, $r(36) = .96$, $p < .01$. All of the Thurstone ability estimates are within 2.0 standard errors of the Rasch ability estimates. The next highest agreement is found between the Rasch and Thorndike ability estimates, $r(37) = .85$, $p < .01$. Thirty-one of the Thorndike ability estimates are within 2.0 standard errors of the Rasch ability estimates. The lowest agreement is found between the Thorndike and Thurstone ability estimates, $r(35) = .78$, $p < .01$.

One reason for the lower correlations obtained with the Thorndike ability estimates and the other two sets of ability estimates is that Thorndike adjusted the ability estimates of some students that he identified as having inconsistent response patterns. Examples of consistent, inconsistent and unscorable response patterns are shown in Table 2. In the sections on inconsistent and unscorable response patterns shown in this table, Thorndike made adjustments in 3 of the

ability estimates (2, 5 & 15). He did not adjust three others (8, 9 & 18) that he identified as inconsistent. Overall, Thorndike adjusted downward the ability estimates of 17 students.

The plots of the Thorndike and Thurstone ability estimates on the Rasch ability estimates are presented in Figures 5 and 6 respectively.

Insert Figures 5 & 6 about here

In Figure 5, it is clear that the Thorndike ability scale consists of 4 to 5 discontinuous levels or steps. This is because Thorndike reported his ability estimates based on the fixed item scale values used to represent the item score groups. Other ability estimates can not be obtained with Thorndike's method. An examination of Figure 6 suggests that the relationship between the Thurstone and Rasch ability estimates may not be linear. The relationships above and below the mean of the Rasch ability estimates are linear, however, a ceiling effect seems to be evident with the Thurstone ability estimates truncated above the mean. This may be due to the use of wrong responses within Thurstone's method of scoring individual performance.

The results of the person fit analyses based on the Rasch model are shown in Table 4. The results for the 11 students identified by Thorndike as having consistent, inconsistent and unscorable response patterns are shown in Table 5. Only one response

Insert Table 5 about here

pattern (student 15) was identified as misfitting (Person fit = 4.07), and this was the same student identified as unscorable by Thorndike. All of the consistent response patterns had non-significant misfit statistics. Only two of the inconsistent response patterns (student 5 & 9) identified by Thorndike had person fit statistics that were somewhat high (Person fit = 1.48).

In order to further illustrate the Rasch person fit analysis, the observed and expected person response curves for student 4 with a consistent response pattern (Person fit = -.08), and student 15 with an inconsistent response pattern (Person fit = 4.07) are shown in Figures 7 and 8 respectively. Both of these students have

Insert Figures 7 & 8 about here

Rasch ability estimates of .571. The mean of the absolute values of the residuals (observed - expected proportion correct) provides a rough index of the impact of person misfit. For student 4 the mean is .12, while the mean is .28 for the student 15.

Within the framework of item-invariant measurement as discussed in this paper, it is also of interest to calculate the ability estimates

for these two students based on the 20 easy and 23 hard vocabulary items which provides an indication of item-invariance. The EASY and HARD ability estimates, as well as standardized differences for selected students are shown in Table 5. For student 4 with a consistent response pattern, the ability estimates on the easy and hard items are 1.265 and .326 respectively. The standardized difference between these two ability estimates is .79. For student 15 with an inconsistent response pattern, the ability estimates on the easy and hard items are -1.660 and 2.038 respectively with a standardized difference of 5.28. It is clear that when the responses are inconsistent (high person fit statistics), then the ability estimates are not invariant over the easy and hard items. Further, even when ability estimates are not available on the easy items (students answered all of the items correctly), the ability estimates on the hard items are quite close to the estimates based on the total number of items when person fit statistics are low.

An attempt was also made to see if the response patterns that led Thorndike to adjust his ability estimates had high person misfit statistics. Other than student 15, who was identified by Thorndike as having an unscorable response pattern, there does not appear to be any relationship between the Rasch person fit statistics and Thorndike's adjustments.

Discussion

The analyses reported in this study suggest that considerable

progress has been made in addressing issues related to item-invariant measurement and the scoring of individual performance. Thorndike, Thurstone and Rasch were working within a common research tradition based on scaling theory. They recognized many of the measurement problems related to invariance, and progress appears in the movement from essentially ad hoc to theory-based measurement procedures for solving these problems. Each of these measurement theorists made significant contributions related to item-invariant measurement, and there are many similarities in the ways in which they conceptualized, modelled and empirically addressed these issues.

It is interesting to note that early measurement theorists, such as Thorndike and Thurstone, generally addressed measurement problems related to invariance as separate activities. This duality between item calibration and the measurement of individuals has appeared repeatedly within the history of psychometrics in the work of Guilford (1936) and Mosier (1940, 1941). In terms of invariance, this has led to a two step process. First, items are calibrated and examined for invariance over subgroups of examinees; second, the invariance of ability estimates over subsets of the calibrated items are examined in some fashion. One of the major breakthroughs in modern measurement theory was the development of item response models, such as the Rasch model, using the individual as the level of analysis which provides for the simultaneous treatment of both item calibration and individual measurement (Lumsden 1976). By

including both ability and item parameters within a probabilistic model, Rasch (1980/1960) proposed a general framework within which these parameters can be estimated separately (Masters & Wright, 1984), and issues related to invariance explicitly addressed.

It should be pointed out that item-invariance is only one part of the problem of obtaining meaningful scores over subsets of items for individuals. In this study, it is assumed that the items are all indicators of an ability related to visual vocabulary. This assumption of unidimensionality is not directly addressed here. Content considerations must also be taken into account when examining ability estimates obtained with different subsets of items. Inconsistencies in response patterns may suggest areas of weakness in student performance, and may also reflect differences in the opportunity to learn the content tested by a particular set of items. Some attention must also be directed towards the precision and accuracy of the item scale values which were assumed to be known for the purposes of this paper.

There are four major points I would like to highlight. First, progress in the scoring of individual performance is reflected in the movement from ad hoc to theory-based methods. Rasch offered a formal statistical model that can be used to represent the empirical person response curves that underlie the methods of both Thorndike and Thurstone. Second, the group-level procedures for examining item-invariance used by these three measurement theorists are quite similar;

each theorist recognized the importance of examining whether or not the relative ordering of individuals is invariant over subsets of items. The graphical methods proposed by Thurstone and Rasch are virtually identical. Third, the recent extensions of the Rasch model for examining the response patterns of each individual provides a probabilistic framework for addressing some of Thorndike's concerns about how to handle randomness and uncertainty within individual response patterns; Thorndike recognized this issue, but offered vaguely defined rules of thumb for dealing with it. Fourth, the data analyzed here suggests that if the response patterns are consistent, then item-invariance has been approximated.

In summary, item-invariant measurement procedures play an important role in the quest for objective measurement in the social sciences. Invariance has been identified as a fundamental aspect of measurement in the behavioral sciences (Bock & Jones, 1968; Engelhard, 1984, 1989; Jones, 1960; Jones & Appelbaum, 1989; Stevens, 1951). When item-invariant measurement is approximated, then the estimation of an individual's ability is not dependent on the particular subset of test items that happen to be used to obtain this estimate.

References

Anderson, E. B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38, 123-140.

Andrich, D. (1988). Rasch models for measurement. Newbury Park: CA: Sage Publications, Inc.

Bock, R. D. & Jones, L. V. (1968). The measurement and prediction of judgement and choice. San Francisco: Holden-Day.

Brown, C. W., Bartelme, P., & Cox, G. M. (1933). The scoring of individual performance on tests scaled according to the theory of absolute scaling. Journal of Educational Psychology, 24, 654-662.

Engelhard, G. (1984). Thorndike, Thurstone and Rasch: A comparison of their methods of scaling psychological tests. Applied Psychological Measurement, 8, 21-38.

Engelhard, G. (1989). Historical views of the concept of invariance and measurement theory in the behavioral sciences. Invited paper presented at the Fifth International Objective Measurement Workshop, University of California, Berkeley.

Gulliksen, H. (1950). Theory of mental tests. New York: Wiley.

Jones, L. V. (1960). Some invariant findings under the method of successive intervals. In H. Gulliksen & S. Messick (Eds.), Psychological Scaling: Theory and applications, (pp. 7-20). New York: John Wiley & Sons, Inc.

Jones, L. V. & Appelbaum, M. I. (1989). Psychometric methods. Annual Review of Psychology, 40, 23-43.

Loevinger, J. (1965). Person and population as psychometric concepts. Psychological Review, 72, 143-155.

Lumsden, J. (1976). Test theory. Annual Review of Psychology, 27, 251-280.

Masters, G. N. & Wright, B. D. (1984). The essential process in a family of measurement models. Psychometrika, 49, 529-544

Mosier, C. I. (1940). Psychophysics and mental test theory: Fundamental postulates and elementary theorems. Psychological Review, 47, 355-366.

Mosier, C. I. (1941). Psychophysics and mental test theory II: The constant process. Psychological Review, 48, 235-249.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. In J. Neyman (Ed.), Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, (pp. 321-333). Berkeley, CA: University of California Press.

Rasch, G. (1966a). An individualistic approach to item analysis. In P. F. Lazarsfeld & N. Henry (Eds.), Readings in Mathematical Social Science (pp. 89-107). Chicago: Science Research Associates.

Rasch, G. (1966b). An item analysis which takes individual differences into account. British Journal of Mathematical and Statistical Psychology, 19, 49-57.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. Danish Yearbook of Philosophy, 14, 58-94.

Rasch, G. (1980/1960). Probabilistic models for some intelligence and attainment tests. Chicago: The University of Chicago Press. [Originally published in 1960 by the Danish Institute for Educational Research].

Smith, R. M. (1985). Validation of individual test response patterns. In T. Husen & T. N. Postlethwaite (Eds.), The International Encyclopedia of Education. Oxford: Pergamon Press.

Smith, R. M. (1986). Person fit in the Rasch model. Educational and Psychological Measurement, 46, 359-372.

Smith, R. M. (1988). The distributional properties of Rasch standardized residuals. Educational and Psychological Measurement, 48, 657-667.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), Handbook of Experimental Psychology, (pp. 1-49). New York: Wiley.

Thorndike, E. L. (1904). An introduction to the theory of mental and social measurements. New York: Teachers College, Columbia University.

Thorndike, E. L. (1910). Handwriting. Teachers College Record, 11, 83-175.

Thorndike, E. L. (1914). The measurement of ability in reading. Teachers College Record, 15, 207-277.

Thorndike, E. L. (1918). The nature, purposes, and general methods of measurements of educational products. In Whipple, G. M. (Ed.), The seventeenth yearbook of the national society for the study of education. Part II, The measurement of educational products. Bloomington, IL: Public School Publishing Company.

Thorndike, E. L. (1921). Measurement in education. Teachers College Record, 22, 371-379.

Thorndike, E.L., Bregman, E. O., Cobb, M. V. & Woodyard, E. (1926). The measurement of intelligence. New York: Bureau of Publications, Teachers College, Columbia University.

Thurstone, L. L. (1925). A method of scaling psychological and educational tests. Journal of Educational Psychology, 15, 433-451.

Thurstone, L. L. (1926). The scoring of individual performance. Journal of Educational Psychology, 17, 446-457.

Thurstone, L. L. (1927). The unit of measurement in educational scales. Journal of Educational Psychology, 18, 505-524.

Thurstone, L. L. (1928a). II. Comment by Professor L. L. Thurstone. Journal of Educational Psychology, 19, 117-124.

Thurstone, L. L. (1928b). Scale construction with weighted observations. Journal of Educational Psychology, 19, 441-453.

Trabin, T. E. & Weiss, D. J. (1983). The person response curve: Fit of individuals to item response theory models. In D. J. Weiss (Ed.), New horizons in testing: Latent trait test theory and computerized adaptive testing (pp.83-108). New York: Academic Press.

Wainer, H. & Wright, B. D. (1980). Robust estimation of ability in the Rasch model. Psychometrika, 45, 373-391.

Wright, B. D. (1968). Sample-free test calibration and person measurement. In Proceedings of the 1967 invitational conference on testing problems. Princeton, NJ: Educational Testing Service.

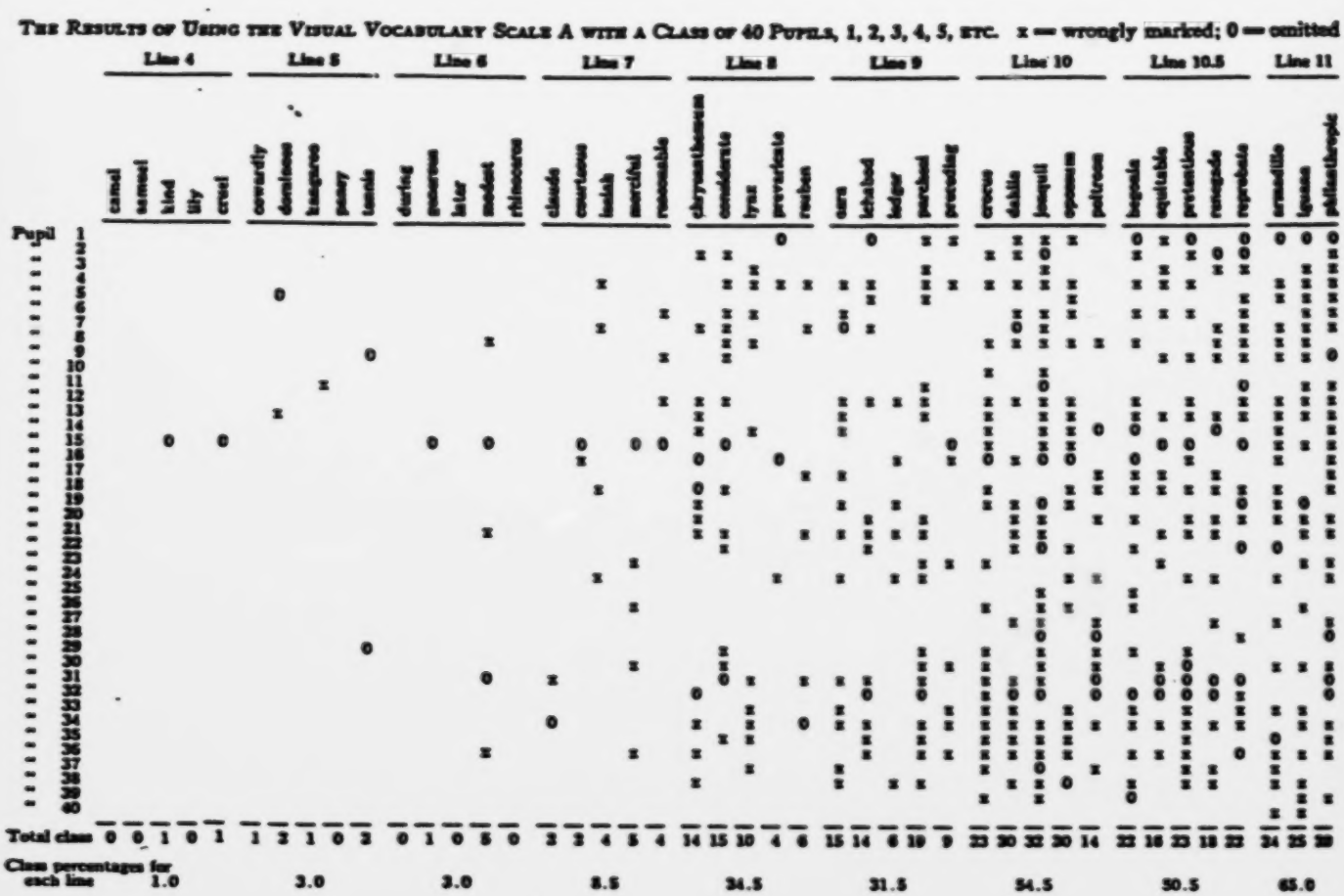
Wright, B. D. & Masters, G. (1982). Rating scale analysis: Rasch Measurement. Chicago: MESA Press.

Wright, B. D. & Panchapakesan, N. (1969). A procedure for sample-free item analysis. Educational and Psychological Measurement, 29, 23-48.

Wright, B. D. & Stone, M. H. (1979). Best test design: Rasch Measurement. Chicago: MESA Press.

Table 1

Individual (Pupil) by Item Matrix



Note. Blank spaces indicate correct responses, and are coded as 1s for the analyses in this paper; wrongly marked (x) and omitted (O) responses are coded as 0s. There are two errors in Thorndike's Table 1: The entry for Pupil 15 for the word "cowardly" should be an x, and the entry for Pupil 13 for the word "dominoes" should be blank. From "The measurement of ability in reading" by E. L. Thorndike, 1914, Teachers College Record, 18, p. 11. Permission to reprint has been requested.

Table 2

Sums of Wrongly Marked and Omitted Words for Selected Student Responses

	Thorndike's Item Scale Values									Ability	
Student	4	5	6	7	8	9	10	10.5	11	T1	T2
Consistent Responses											
1	0	0	0	0	1	3	3	4	3	1.704	1.704
3	0	0	0	0	1	1	1	3	2	3.704	3.704
12	0	0	0	1	2	4	4	3	3	.704	.704
13	0	0	0	0	1	2	3	5	3	1.704	1.704
25	0	0	0	0	0	0	1	1	0	4.704	4.704
Inconsistent Responses											
2	0	0	0	0	2	0	3	4	1	2.704	4.704*
5	0	1	0	0	0	2	1	1	3	3.704	4.204*
8	0	0	1	0	2	0	5	3	2	2.704	2.704
9	0	1	0	1	1	0	0	4	3	3.704	3.704
18	0	0	0	1	2	0	3	5	2	2.704	2.704
Unscorable Response											
15	2	1	2	3	1	1	3	3	3	?	2.704*

Note. There are 5 items for each item scale value except the last (11) which has 3 items. The T1 ability estimates are taken from Thorndike (1914, p. 12), and reported on an adjusted scale. The T2 ability estimates are based on the strict application of Thorndike's scoring rules; T2 ability estimates with asterisks were adjusted by Thorndike.

Table 3

Comparisons of Thorndike (THORN), Thurstone (THURS) and Rasch (RASCH) on Major Issues

Issue	THORN	THURS	RASCH
Recognized importance of item-invariant measurement	Yes	Yes	Yes
Utilized concept of latent variable scale	Yes	Yes	Yes
Avoided using raw scores	Yes	Yes	Yes
Person response curves	Yes	Yes	Yes
Formal probabilistic model	No	No	Yes
Standard errors for ability estimates	No	No	Yes
Scoring criterion	80%	50%	50%
Flagged inconsistent response patterns	Yes (ad hoc)	No	Yes (theory)
Item calibration	Separate Process	Separate Process	Simultaneous Process

Table 4

Ability Estimates Based on Thorndike (THORN), Thurstone (THURS) and Rasch (RASCH) Methods with Rasch Standard Errors (SE) and Person Fit Statistics

Student	THORN	THURS	RASCH	RASCH SE	Person Fit
1	1.704	2.114	1.545	0.441	-0.69
2	2.704	2.614	2.338	0.453	-0.20
3	3.704	2.814	2.764	0.471	-0.22
4	0.704	1.114	0.571	0.443	-0.08
5	3.704	2.814	2.764	0.471	1.48
6	0.704	2.314	1.740	0.442	-0.08
7	0.704	2.314	1.740	0.442	0.08
8	2.704	2.314	1.740	0.442	0.34
9	3.704	2.614	2.338	0.453	1.48
10	4.704	-	4.665	0.759	-0.13
11	4.204	3.014	3.238	0.506	1.58
12	0.704	1.514	0.962	0.441	-0.32
13	1.704	1.914	1.350	0.441	0.76
14	1.704	2.714	2.135	0.448	-0.22
15	?	1.114	0.571	0.443	4.07
16	0.704	2.314	1.740	0.442	0.04
17	3.704	2.914	2.992	0.486	-0.12
18	2.704	2.314	1.740	0.442	-0.17
19	1.704	2.614	2.338	0.453	-0.24
20	1.704	2.314	1.740	0.442	-0.65

Table 4 (Cont.)

Student	THORN	THURS	RASCH	RASCH SE	Person Fit
21	0.704	1.914	1.350	0.441	0.47
22	2.704	2.814	2.764	0.471	-0.16
23	3.704	2.914	2.992	0.486	0.22
24	1.704	2.514	2.135	0.448	0.09
25	4.704	-	4.665	0.759	-0.17
26	3.704	3.014	3.238	0.506	0.23
27	2.704	3.014	3.238	0.506	-0.39
28	3.704	3.214	3.814	0.576	-0.28
29	2.704	2.814	2.764	0.471	1.54
30	1.704	2.514	1.936	0.445	-0.08
31	0.704	1.514	0.962	0.441	0.67
32	1.704	2.314	1.740	0.442	-0.46
33	1.704	2.714	2.135	0.448	-0.35
34	0.704	0.914	0.374	0.444	-0.39
35	0.704	2.514	2.135	0.448	-0.15
36	1.704	1.514	0.962	0.441	-0.05
37	2.704	2.814	2.764	0.471	-0.16
38	1.704	2.514	1.936	0.445	-0.36
39	2.704	3.114	3.507	0.534	-0.36
40	4.204	3.414	4.665	0.759	-0.28
M	2.319	2.559	2.277	0.484	.16
SD	1.264	.741	1.076	0.085	.85

Table 5

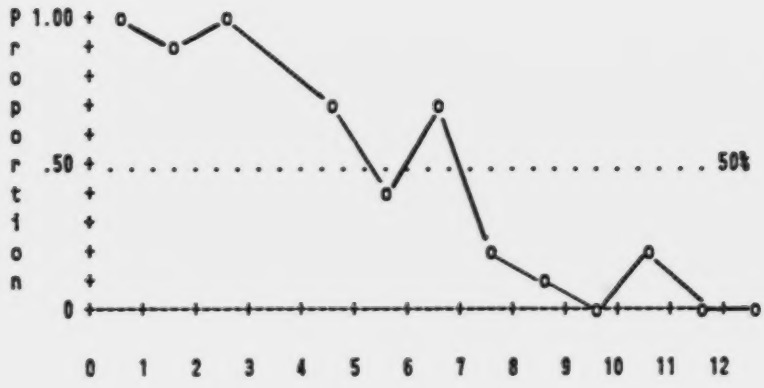
Ability Estimates Based on Thorndike (THORN), Thurstone (THURS) and Rasch (RASCH) Methods for Students with Consistent, Inconsistent and Unscorable Responses Patterns

Student	THORN	THURS	RASCH	RASCH (SE)	Person Fit	RASCH EASY	RASCH HARD	Standardized Difference
Consistent Responses								
1	1.704	2.114	1.545	0.441	-0.69	-	1.368	-
3	3.704	2.814	2.764	0.471	-0.22	-	2.708	-
12	0.704	1.514	0.961	0.441	-0.32	1.265	.882	-.33
13	1.704	1.914	1.350	0.441	0.76	1.265	1.368	.09
25	4.704	-	4.665	0.759	-0.17	-	4.644	-
Inconsistent Responses								
2	2.704	2.614	2.338	0.453	-0.20	-	2.258	-
5	3.704	2.814	2.764	0.471	1.48	1.265	2.945	1.45
8	2.704	2.314	1.740	0.442	0.34	1.265	1.818	.48
9	3.704	2.614	2.338	0.453	1.48	.468	2.708	2.43
18	2.704	2.314	1.740	0.442	-0.17	1.265	1.818	.48
Unscorable Response								
15	?	1.114	0.571	0.443	4.07	-1.660	2.038	5.28

Note. Dashes indicate that ability estimates are not obtainable. The EASY and Hard ability estimates for the Rasch model are based on the 20 easy and 23 hard items respectively. The standardized difference is the HARD minus the EASY ability estimates divided by a pooled estimate of the standard error of the difference.

Figure 1

Thurstone's Empirical Person Response Curve

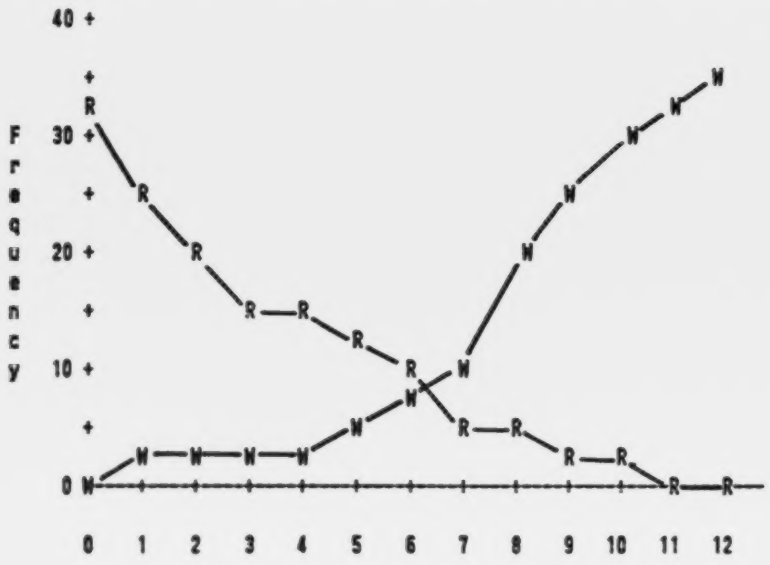


Right	6	9	4	0	2	2	6	1	1	0	1	0	0
Wrong	0	1	0	0	1	3	3	3	7	7	5	2	4
Total	6	10	4	0	3	5	9	4	8	7	6	2	4
Proportion correct													
	1.00	.90	1.00	-	.67	.40	.67	.25	.12	.00	.17	.00	.00

Note. Proportion correct for items grouped by the item scale values (0 to 12) for one individual. Adapted from Thurstone (1926, p. 449).

Figure 2

Thurstone's Method of Scoring Individual Performance



Cumulative Frequency Right (R)

32 28 17 13 13 11 9 3 2 1 1 0 0

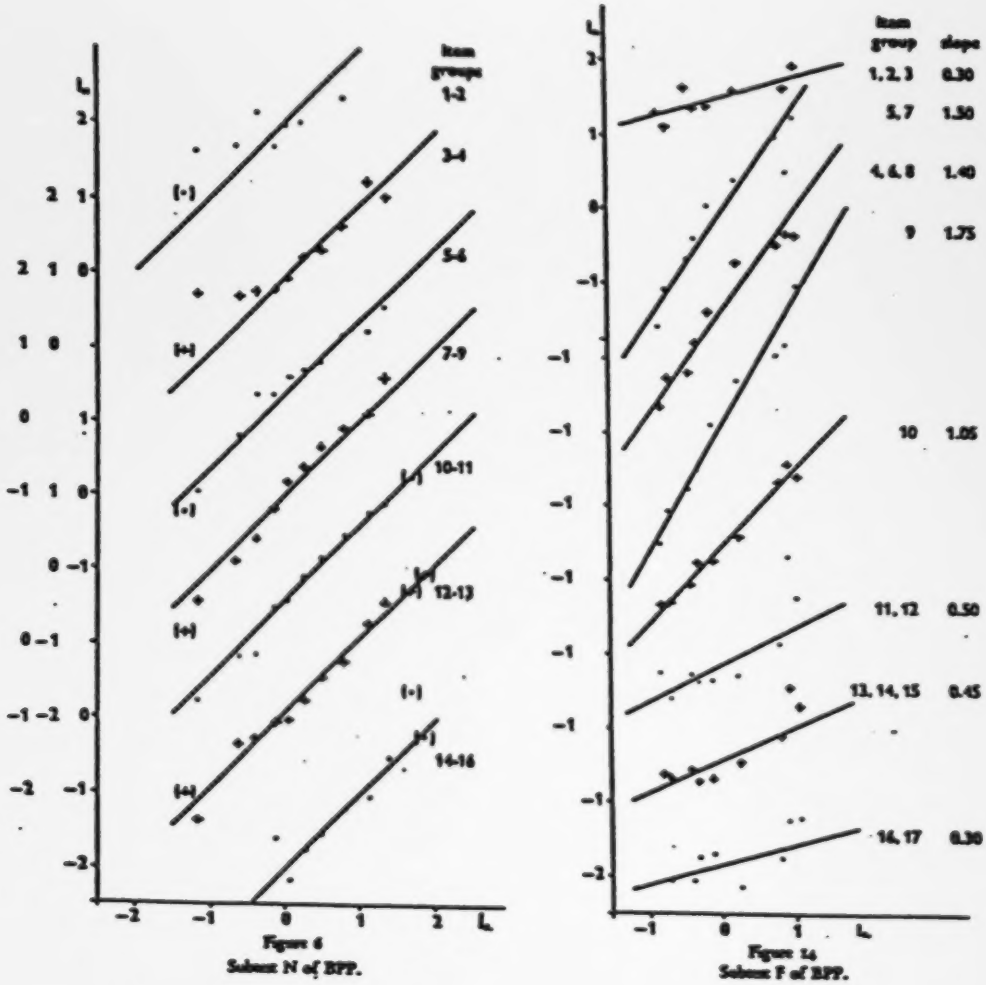
Cumulative Frequency Wrong (W)

0 1 1 1 2 5 8 11 18 25 30 32 36

Note. Cumulative Frequency of right (R) and wrong (W) responses for items grouped by the item scale values (0 to 12) for one individual. Based on the same data presented in Figure 1. Adapted from Thurstone (1926, p. 451).

Figure 3

Rasch's Graphic Approach for Examining the Item-Invariant Measurement



A. Successful item-invariant measurement

B. Unsuccessful item-invariant measurement

Note. The abscissa (lr.) in each panel is the average of the ability estimates calculated separately within item groups. The ordinate (lri) represents the ability estimates calculated within each item group with a constant added by Rasch to avoid overlapping ability estimates and to highlight the linearity or non-linearity of these plots. From Probabilistic models for some intelligence and attainment tests (pp. 87 & 97) by G. Rasch, 1960/1980, Chicago: The University of Chicago Press. Copyright 1960 by The University of Chicago. Permission to reprint has been requested.

Figure 4

Expected Person Response Curve Based on the Rasch Model

(Ability = .571)

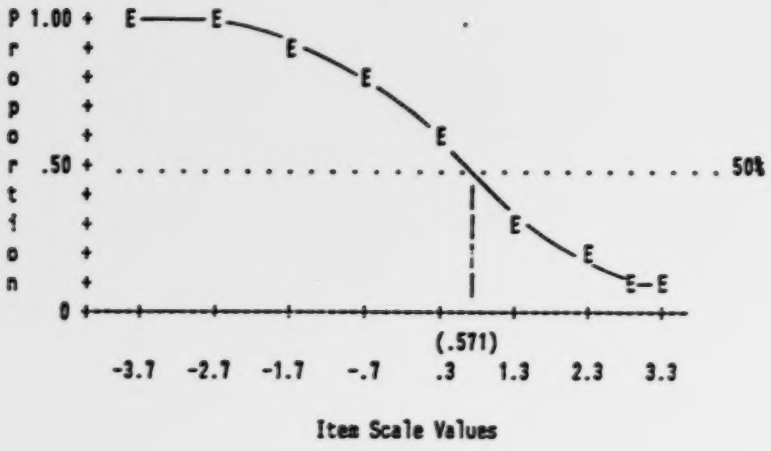


Figure 5

Plot of Thorndike (THORN) and Rasch (RASCH) Ability Estimates

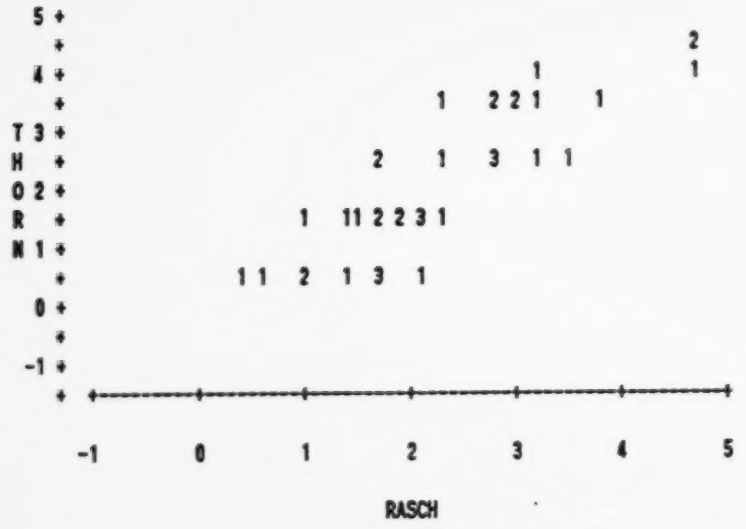


Figure 6

Plot of Thorndike (THURS) and Rasch (RASCH) Ability Estimates

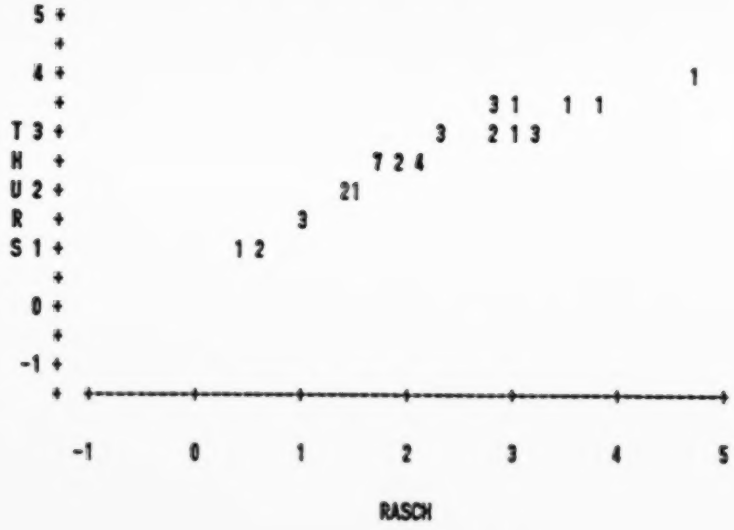
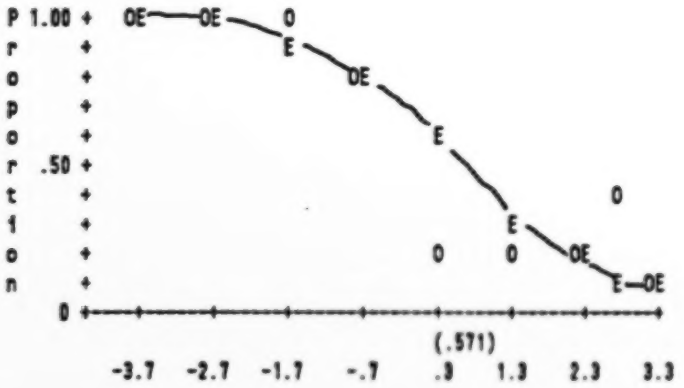


Figure 7

Observed and Expected Person Response Curves Based on the Rasch Model for Student 4 with a Consistent Response Pattern (Ability = .571, Person fit = -.08).

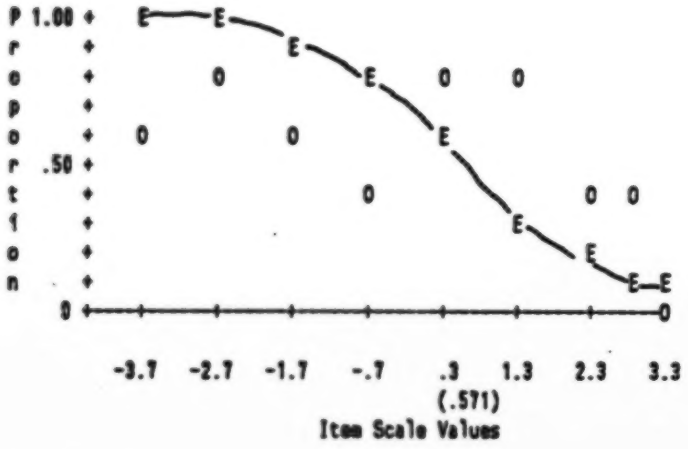


Item Score Groups

	-3.7	-2.7	-1.7	-.7	.3	1.3	2.3	2.8	3.3
Observed (O)	1.000	1.000	1.000	.800	.200	.200	.200	.400	.000
Expected (E)	.986	.963	.905	.778	.564	.322	.149	.096	.060
Residual (O - E)	.014	.037	.095	.022	-.364	-.122	.051	.304	-.060

Figure 8

Observed and Expected Person Response Curves Based on the Rasch Model for Student 15 with an Inconsistent Response Pattern (Ability = .571, Person fit = 4.07).



Item Score Groups

	-3.7	-2.7	-1.7	-.7	.3	1.3	2.3	2.8	3.3
Observed (O)	.600	.800	.600	.400	.800	.800	.400	.400	.000
Expected (E)	.986	.963	.905	.778	.564	.322	.149	.096	.060
Residual (O - E)	-.386	-.163	-.305	-.378	.236	.478	.251	.304	-.060